10-2016

Attractiveness versus competition: Towards an unified model for user visitation

Thanh-Nam DOAN
Singapore Management University, tndoan.2012@phdis.smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1145/2983323.2983657

Attractiveness versus Competition: Towards an Unified Model for User Visitation

Thanh-Nam Doan
School of Information Systems
Singapore Management University
tndoan.2012@phdis.smu.edu.sg

Ee-Peng Lim
School of Information Systems
Singapore Management University
eplim@smu.edu.sg

ABSTRACT

Modeling user check-in behavior provides useful insights about venues as well as the users visiting them. These insights can be used in urban planning and recommender system applications. Unlike previous works that focus on modeling distance effect on user's choice of check-in venues, this paper studies check-in behaviors affected by two venue-related factors, namely, *area attractiveness* and *neighborhood competitiveness*. The former refers to the ability of an area with multiple venues to collectively attract check-ins from users, while the latter represents the ability of a venue to compete with its neighbors in the same area for check-ins. We first embark on a data science study to ascertain the two factors using two Foursquare datasets gathered from users and venues in Singapore and Jakarta, two major cities in Asia. We then propose the VAN model incorporating user-venue distance, area attractiveness and neighborhood competitiveness factors. The results from real datasets show that VAN model outperforms the various baselines in two tasks: home location prediction and check-in prediction.

Keywords

Neighborhood Competition, Area attractiveness, location-based social network

1. INTRODUCTION

Motivation. The popularity of smartphones and wearable devices in recent years has helped to create new location based social networking (LBSN) applications for users to publish their visits (or check-ins) to different venues. By analyzing these check-in data, one may derive useful insights for urban planning, business recommendation, and other applications.

Previous works on LBSN data have shown that users prefer to visit venues near their home locations[7, 5]. This is also known as the *distance effect*. It underscores the importance of home location of users when analyzing their movement. Other than the distance effect which is specific to a user-venue pair, there are other venue factors that have not yet been studied and modeled.

In this paper, we introduce *area attractiveness* and *neighborhood competitiveness* as two new venue factors for analyzing and modeling check-in behavior. Area attractiveness refers to the ability of an area with multiple venues to collectively attract check-ins from users. Neighborhood competitiveness specifies the ability of a venue to compete with its neighbors in the same area for check-ins. We hypothesize that when a user decides a venue to visit, she will first select an area before finalizing the venue in the area. This two stage process suggests that some areas attract more visitors than others. The choice of area will reduce the cognitive load on the user as she has fewer candidate venues in the area to choose from.

Learning the area attractiveness and neighborhood competitiveness factors from check-in data gives rise to several useful applications. Urban planners can redesign a city's transportation network by making attractive areas more accessible. Businesses need to know both area attractiveness and neighborhood competitiveness in order to decide the new store locations. A personalized store recommendation app can also leverage on the two factors when making suggestions to its users.

Research Objectives. In this research, we therefore aim to incorporate area attractiveness and neighborhood competition into a new unified model for analyzing user-venue check-in behavior. This model should also include the distance effect which has already been used in the earlier models.

There are however several research challenges. Firstly, area attractiveness and neighborhood competitiveness are new concepts that have not been formally studied earlier. It is not easy to illustrate the effects of these two factors using real data. Hence, there is a need to conduct data science research on the factors. Secondly, the check-ins from users to venues are the results of multiple user and venue factors interacting with one another. Exactly how the interaction takes place is unclear. The challenge is therefore to create some generative stories to describe this interaction. Finally, there is no obvious ground truth in real datasets for evaluation of proposed models. We will need to adopt an indirect approach to conduct model evaluation.

Our results and findings of this research are as follows:

- We carefully gather Foursquare check-in data of users and venues from two cities, Singapore and Jakarta. Next, we determine the exact home locations of a subset of users through some stringent criteria. This gives us good datasets to embark on this research.
- We conduct an empirical analysis of the gathered check-in data and demonstrate the existence of neighborhood competitiveness, area attractiveness factors and also distance effect.
- We propose a probabilistic model called **VAN** to capture the check-in behavior of users incorporating these above effects.
- The performance of our proposed model is evaluated on real

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983657

datasets so as to demonstrate its superior accuracy. Specifically, we apply our proposed model and other baselines to two crucial tasks: home location prediction and check-in prediction. We show that our proposed model outperforms the baselines with reasonable results.

2. RELATED WORKS

Check-in prediction research: Chen *et al.* [4] surveyed the performance of matrix factorization in check-in prediction. Then, they proposed a model to combine matrix factorization and multi-center Gaussian model to predict the check-ins in LBSN with social information as the regularization. However, the work does not consider distance from users to areas and the neighborhood competition among venues. Cho *et al.* [5] viewed check-in locations of users as the mixture of check-ins near *home* and *work*. Our differences are that we assume only one home location for each user and the size of area in our model is predefined as parameter.

Home location identification: There are some works [10, 1] tackling home location identification problem but they predict the home locations of users at city level instead of giving exact locations.

Neighborhood Competition: Hu *et al.* [8] showed that there is weak correlation in rating between venue and its neighbors. Moreover, this correlation is *independent* of venue category. However, their model does not give the exact home location of users nor the ranking of venues. Doan *et al.* [6] proposed a new model named *Competitive Rank* based on *PageRank* to mine the competition between venues and their neighbors. The drawback of this model is that it considers every check-in to be the same but ignores the distance from the user to the check-in venue. To the best of our knowledge, the Huff model [9] measures the competitiveness of a store by its size but size information is not available in practice. Qu *et al.* [13] incorporates information from social network into Huff model. This extended model also replace users' home location by users' activity centers but not using the home location of users can reduce the performance in check-in prediction[7].

3. EMPIRICAL ANALYSIS OF CHECK-IN BEHAVIORAL DATA

In this section, we describe the two datasets as well as preprocessing methods to clean the data. We then use them to conduct empirical analysis on the check-in behavior to find the evidence of *distance effect*, *area attractiveness* and *neighborhood competition*.

3.1 Datasets

To study user check-in behavior and how it is affected by user and venue characteristics, we need Foursquare datasets that capture complete check-in data of both users and venues. We thus decided to crawl check-in data of users and venues in two cities, Singapore and Jakarta. Due to Foursquare's crawling restrictions, we could only crawl the publicly visible check-in data via Twitter APIs.

SG Dataset. This dataset consists of 1.11 millions check-ins by 55,891 Singapore Foursquare users on 75,346 venues between August 15, 2012 and June 3, 2013 (see Table 1). The users and venues are determined to be located in Singapore based on their profile declared location and venue's geo-location respectively.

JK Dataset. Similarly, we crawled another Foursquare dataset for the users and venues in Jakarta the largest city in Indonesia from July 2014 to May 2015. There are 119,618 check-ins performed by 14,974 users on 38,183 venues. The numbers are generally smaller than those of **SG** dataset.

Users with home locations. Among the users in the **SG** (or **JK**) datasets, we identify a subset of users whose home locations can be determined using the method described in Section 3.2. We then construct another dataset to include users with home locations and the venues that they perform check-ins on. This leads us to the **H_SG** and **H_JK** datasets. The number of users with home location in **H_SG** and **H_JK** are 856 and 455 respectively. Compared with **SG** and **JK** users, the users in **H_SG** and **H_JK** are relatively more active in performing check-ins. For example, **SG** has about 20 check-ins per user while **H_SG** has about 74 check-ins per user.

Table 1: Dataset Statistics

	SG	H_SG	JK	H_JK
# users	55,891	856	14,974	455
# venues	75,346	12,020	38,183	4,380
# check-in's	1.11M	63,777	119,618	9,557
# user-venue pairs with > 0 check-ins	541,588	28,298	81,188	5,422

3.2 Home Location Identification

We selected a subset of users whose home locations can be clearly identified using both their check-ins and check-in messages. The following are the detailed steps:

- We selected a subset of venues under the "home (private)" category which is in turn a sub-category of the "residence" category. We found 8447 and 1985 venues satisfying this criteria in the **SG** and **JK** datasets respectively.
- We further identified 3276 and 891 users who performed check-ins at only one "home (private)" venue each in the **SG** and **JK** datasets respectively. This rules out users who performed check-ins at multiple "home (private)" venues.
- We finally selected an even smaller set of users who also shouted some home relevant messages during their check-ins to their "home (private)" venues. These messages have to include some "home" related key phrases, e.g., "back home", "home finally", etc.. For the **JK** dataset, we use the matching Malay key phrases like "Tidur dulu" (sleep first), "Rumah" (House), "Pondok"(cottage), "sampai di rumah" (arrived to home), "bobo"(sleep).

We finally obtained 856 users with home locations among (1.5% of all **SG** users) those users in the **SG** dataset. We denote the Foursquare dataset of these users and their check-in venues by **H_SG**. These users have 63,777 check-ins on 12,020 venues as shown in Table 1. Note that this represents 1.5% of all users and 5.7% of all check-ins in **SG**. Similarly, we obtained the **H_JK** dataset for 455 Jakarta users (3% of all **JK** users) with home locations. These dataset covers 4380 venues and 9557 check-ins.

3.3 Distance Effect

Using **H_SG** and **H_JK** datasets with exact users' home locations, we are able to study distance effect within a city, i.e., distance between the geo-coordinates of user's home location and venue location. We derive the probability of users performing check-ins on venues that are within a distance bin away from users' reported home locations. Specifically, for each user, we divide the city into several circular rings with the user's home location at the center. Every ring has a width of 1 kilometer. In other words, the first ring covers distance range [0,1km), the second ring covers distance range [1km, 2km), and so on. As the large distance rings involve the check-ins of very few users, we exclude rings with distance larger than 25 km. Next, we compute the probability of the user performing check-ins on venues within each distance ring. We finally compute for each distance ring the average probability of all users performing check-ins on venues within that distance ring.

We plot the average probability for different distance rings in Figure 1(a) and observe that: (a) users are more likely to visit or check into venues nearer to their home locations; (b) the decreasing probability trend appears in both **H_SG** and **H_JK** datasets; (c) the probability becomes more stable for venues within the distance rings 10 to 15 km away from users in both datasets.

3.4 Area Attractiveness

Despite the distance effect, some venues may still attract check-ins from users far away. Li *et. al.* [10] developed an influence scope model for measuring the attractiveness of venues to their followers. In this paper, instead of examining attractiveness at the venue level, we model attractiveness at the area level. There are three significant advantages of doing so. Firstly, it reduces the number of parameters in modeling which in turn reduces the learning time. Secondly, we address data sparsity issue at the venue level. Finally, we believe that the area a venue belongs to has a major influence over its ability to attract users. Due to space limitations, we are going to illustrate this by the following empirical analysis on only **H_SG** dataset.

We empirically select three well known fast food chains, i.e., McDonald, KFC and Starbucks, with many branches. We expect branches of the same chain to be very similar to one another by food variety, food quality, ambience and service. Hence, at the venue level, we should not expect any difference among their abilities to attract users from other locations. We now divide the city into square areas of width equals to 0.05 degree (equivalent to about 5.55 km on the equator) and assign every venue to exactly one area. The location of each area is its center of the mass derived from the locations of its venues. The detail of area construction is in Section 4. We call the top five areas with most number of venues the *dense areas* while the areas from ranks 10 to 15 the *sparse areas*. We exclude other lower ranked areas as they do not contain any of the three fast food venues.

For each fast food chain, we examine the distances between each dense area (represented by its center of mass) and the home locations of users who perform check-ins to its venues inside the area. We then generate a boxplot for the user-area distance of all dense areas. We perform the same procedure for sparse areas. Figure 1(b) shows that for each fast food chain, branches within the *dense areas* attract users farther than branches in the *sparse areas*. This suggests that the attractiveness of area plays an important role bringing far away users to the venues in the area.

3.5 Neighborhood Competition Effect

To show competition among venues within the same area, we adopt the method originally proposed by Weng *et al.*[14] to study competition among memes. We divide the check-in history into weeks. We then measure the following entropies for each week.

- **System entropy (E_s):** $E_s(t) = -\sum_v f_v(t) \log f_v(t)$ where $f_v(t)$ is the fraction of check-ins in week t performed on venue v, i.e., $f_v(t) = \frac{\#cks(v,t)}{\sum_v \#cks(v,t)}$. The system entropy essentially measures the degree to which the distribution of check-ins concentrates on a small fraction of venues.

- **Average area entropy (E_A):** We first define the entropy of an area a to be $E_a(t) = -\sum_{v \in a} f_{v,a}(t) \log f_{v,a}(t)$ and $f_{v,a}(t) = \frac{\#cks(v,t)}{\sum_{v \in a} \#cks(v,t)}$. We then take the average of all area entropies, i.e., $E_A(t) = Avg_a E_a(t)$. We divide the city into square cells of 0.05 degree width. The construction of areas is discussed further in Section 4. Similar to system entropy, average area entropy captures the degree to which the distribution of check-ins of an area concentrates on a small fraction of venues (in the area).

- **Average user entropy (E_U):** We next define the average user entropy as $E_U(t) = Avg_{u \in U} E_u(t)$ where entropy of user u is $E_u(t) = -\sum_v f_{u,v}(t) \log f_{u,v}(t)$ and $f_{u,v}(t) = \frac{\#cks(u,v,t)}{\#cks(u,t)}$. This entropy quantifies the concentration of users' attention on the venues they perform check-ins on.

Figure 1(c) shows the three entropies over weeks in both **H_SG** and **H_JK** datasets. The first important observation is that the average user entropy is much smaller than system entropy. It clearly suggests that each user's attention is limited to very small fraction of venues in the entire city. Venues therefore have to compete to gain attraction from users. Secondly, we observed from Figure 1(c) that system entropy is much larger than average area entropy in both datasets. This implies that check-ins within an area concentrated on smaller fraction of venues than the fraction of venues in the entire city receiving check-ins from the whole user population.

The above empirical analysis concludes that venues compete more with their nearby neighbors than those farther away. Thus, grouping venues into areas and modeling competition among venues in each area is an appropriate modeling approach.

4. VISITATION BY ATTRACTIVENESS AND NEIGHBORHOOD COMPETITION MODEL

In this section, we propose the ***Visitation by Attractiveness and Neighborhood competition*** (VAN) model and study its parameter learning process.

4.1 Model Description

Let U and V denote the set of users and venues in a city respectively. We divide the city into mutually exclusive square cells of width s. We use a_v to denote the square or *area* which contains v. More notations and their meanings are shown in Table 2.

The **VAN** model makes the following assumptions for each check-in between user and venue:

- First of all, every user chooses an area to perform a check-in based on its attractiveness and the distance between the user and the area.

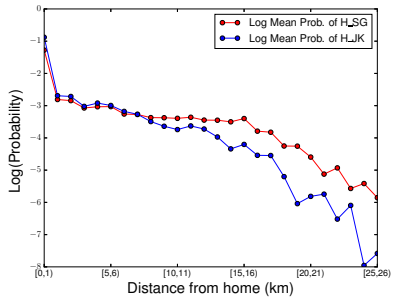
- Secondly, every venue must compete against their neighboring venues in order to gain a check-in.

We assign each venue v a **competitiveness** value σ_v to measure its ability to compete with its neighbors. The value of σ_v is positive, and the larger the σ_v the more competitive the venue v.

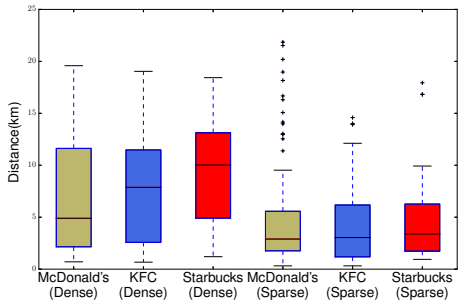
The **neighbors** of a venue v, $N(v)$, are venues within a_v and the areas adjacent to a_v denoted by $Adj(a_v)$. That is, $N(v) = \{v' | v' \in Adj(a_v)\} \cup \{v' | v' \in a_v\} \setminus \{v\}$. We consider the venues in $Adj(a_v)$ as neighbors because we want to include venues in these nearby area as competitors of v. Otherwise, the competing neighbors of some venues near the border of a_v might not be included.

The **center** of area a is defined as the center of the mass location of venues inside a. The **attractiveness** of area σ_a is defined by the root mean square of the **competitiveness** scores of venues in a. That is, $\sigma_a = \sqrt{\sum_{v \in a} \sigma_v^2}$. It means that the venues inside the area contribute their competitiveness together to attract user check-ins.

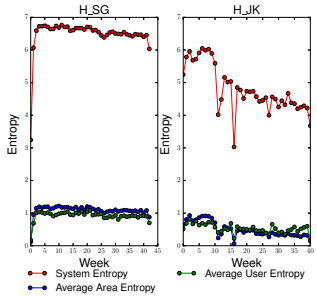
Every check-in of user i to venue v follows a two-step process. Firstly, user i must select the area a_v. Secondly, the venue v in area a_v must win over all other neighboring venues in $N(v)$ to gain a check-in from user i.



(a) Fraction of check-ins as a function of distance from home in **H_SG** and **H_JK** datasets in log scale.



(b) Boxplot of distance from areas containing fast food chain to their check-ins users in **H_SG**



(c) Weakly entropy in **H_SG** and **H_JK** datasets.

Figure 1: The plots of empirical studies.

Table 2: Table of Notations.

Notations	Meaning
U / V / C	set of all users/venues/check-ins
w_{iv}	number of check-in of user i to venue v
w_v	total number of check-in of venue v
a_v	area a_v containing venue v
s	the width of area
σ_v	competitiveness score of venue v
σ_{a_v}	attractiveness of area a_v
$N(v)$	set of neighbor venues of v
$CDF(\cdot)$	cumulative density function of standard normal distribution
$S(\cdot)$	Sigmoid function
$p(i \rightarrow a_v)$	probability of user i visiting area a_v

- User i selects the area a_v under the effect of attractiveness of area a_v. Moreover, if the distance between i and a_v increases, the probability of user i choose area a_v decreases. We model this by zero-mean Gaussian distribution whose variance is σ_{a_v}. The Euclidean distance between user i and a_v is the random variable generated from the distribution. In other words, the home location of user i is generated from the Gaussian distance whose mean is the location of area a_v and variance is σ_{a_v}.

- To model the winning of venue v over its neighbors, we need to model the difference of competitiveness of v and that of one of its neighbor, say v' . We propose two options: *cumulative distribution function (CDF)* of standard Gaussian distribution i.e. $CDF(\sigma_v - \sigma_{v'}; 0, 1)$ and *Sigmoid function* of $\sigma_v - \sigma_{v'}$, i.e. $S(\sigma_v - \sigma_{v'})$. Both functions map differences between the competitiveness values of two venues into the range $[0, 1]$. If venue v is more competitive than its neighbor v' i.e. $\sigma_v > \sigma_{v'}$, the two functions will return a higher probability of v winning the check-in over v'.

Example: Figure 2 depicts two check-ins at venue v by user i i.e. $w_{iv} = 2$. To perform each check-in at venue v, user i has to select area (b, 3)(enclosed by red box) considering the distance from his home location to the center of area (b, 3) and the attractiveness of area (b, 3). Moreover, the venue v needs to *win* over all of its neighbors in the adjacent areas (i.e. venues within the green box).

4.2 Formalization & Inference

The probability p_{iv} of a check-in from user i to venue v is:

$$p_{iv} = p(i \rightarrow a_v) \prod_{v' \in N(v)} p(v > v') \tag{1}$$

Equation 1 says that p_{iv} depends on two components: $p(i \rightarrow a_v)$ denoting the probability of user i selecting area a_v and $p(v > v')$ denoting the probability of venue v winning over its neighbor v'.

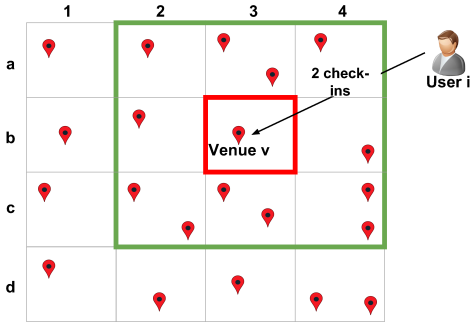


Figure 2: Example of Check-in graph.

Let (x_i, y_i) and (x_{a_v}, y_{a_v}) denote the location of user i and center of area a_v respectively. The probability $p(i \rightarrow a_v)$ is defined:

$$p(i \rightarrow a_v) = \mathcal{N}((x_i, y_i); (x_{a_v}, y_{a_v}), \sigma_{a_v}^2) \tag{2}$$

We model the attractiveness of each area a_v by a bivariate Gaussian distribution with center of area as mean and covariance matrix representing the attractiveness of a_v, i.e., σ_{a_v}. The larger Euclidean distance between user i and center of area a_v, the smaller the $p(i \rightarrow a_v)$. The covariance matrix is diagonal and has same value σ_{a_v} because we assume that the attractiveness of area a_v in *x-axis* is similar to its attractiveness in *y-axis*.

The log-likelihood of a set of check-ins C from users from U on venues from V is then defined.

$$\begin{aligned}
\mathcal{L}(C|\{\sigma_v\}_{v \in V}) &= \sum_{(i,v) \in C} w_{iv} \log p_{iv} \\
&= \sum_{(i,v) \in C} w_{iv} \log p(i \rightarrow a_v) + \sum_v w_v \sum_{v' \in N(v)} \log p(v > v') \\
&= \sum_{(i,v) \in C} w_{iv} \left(-2 \log \sigma_{a_v} - \frac{1}{2\sigma_{a_v}^2} ((x_i - x_{a_v})^2 + (y_i - y_{a_v})^2) \right) \\
&\quad + \sum_v w_v \sum_{v' \in N(v)} \log p(v > v') + const
\end{aligned} \tag{3}$$

As shown in Table 2, w_{iv} denotes the number of check-ins between user i and venue v, and w_v denotes the total number of check-ins on venue v. We consider two options to model the probability $p(v > v')$, i.e., Sigmoid function and cumulative density function of standard Gaussian distribution. Depending on the choice of the above options, we derive two variants of VAN models denoted by $\textbf{VAN}_{Sigmoid}$ and $\textbf{VAN}_{CDF}$.

Inference of user home locations: Taking derivative *w.r.t.* the

x-coordinate of user i's home location and set it to 0 gives us:

$$\frac{\partial \mathcal{L}}{\partial x_i} = \sum_v w_{iv}\left(-\frac{1}{2\sigma_{a_v}^2}2(x_i - x_{a_v})\right) = 0$$
$$x_i = \frac{\sum_v \frac{w_{iv}}{\sigma_{a_v}^2} x_{a_v}}{\sum_v \frac{w_{iv}}{\sigma_{a_v}^2}} \text{ and } y_i = \frac{\sum_v \frac{w_{iv}}{\sigma_{a_v}^2} y_{a_v}}{\sum_v \frac{w_{iv}}{\sigma_{a_v}^2}} \quad (4)$$

Based on Equations 4, we derive some interesting observations about the home location of user i.

- The home location of user i is the *weighted average* of centers of area of venues checked in by i.
- The weight associated to each area has two components: the number of check-ins of user i to venue in the area and the attractiveness of the area. The former helps to predict the home location close to the check-in area due to distance effect. However, area attractiveness has an inverse effect on the importance of area. That is, more attractive areas should contribute less to identifying the home location of user i.
- Suppose the maximum and minimum of x-coordinate (i.e. latitude) of city are x_{max} and x_{min} respectively so we have $\forall a_v : x_{min} \leq x_{a_v} \leq x_{max}$. Thus, $\forall i \in U : x_{min} \leq x_i \leq x_{max}$. Similarly, $\forall i \in U : y_{min} \leq y_i \leq y_{max}$ for y-coordinate. Hence, the weighted average of centers of check-in areas ensures that the home location of user i to be within the city boundary.

Inference of competitiveness of venues: To maximize $\mathcal{L}$ with respect to σ_v and the constraint $\sigma_v > 0$, we add the regularization term $\sum_{v\in V} \log \sigma_v$ and use *gradient descent* with *back-tracking*[3] to find the optimal values of σ_v. The regularization term $\sum_{v\in V} \log \sigma_v$ helps us to keep all σ_v positive because if $\exists v \in V : \sigma_v \to 0$, then $\log \sigma_v$ and $\sum_{v\in V} \log \sigma_v$ will become $-\infty$.

5. EVALUATION USING REAL DATA

We evaluate the VAN models on real dataset in two separate tasks: home location prediction task and check-in prediction task. The geography degree is chosen as the unit of parameter s.

5.1 Home location prediction

Description: In this task, we aim to predict the home locations of users using our VAN models and some baselines. Among the baseline methods for comparison, Periodic Mobility Model (PMM) [5] is the state-of-the-art home location prediction method.

Setup: In total, we have the exact home location of 856 users in **H_SG** dataset. However, there are 341 of them whose home location cannot be predicted by PMM model as these users have too few check-ins or too few venues not giving PMM enough data to learn their home locations. Hence, we will conduct the experiment on the remaining 515 users. In the experiment, we randomly separate 515 users into five equal folds. For each run, we hide the home location of users in one fold and use all check-in data from all five folds and home location of users from the remaining four folds as input. Each model will then predict the home location of users in the hidden fold. For PMM, only the check-in data of users is used to predict their home locations. Hence, each time, we select one fold and predict home location of users in that fold by their check-in data. Similar to **H_SG**, there are 154 out of 455 users in **H_JK** whose home locations could be predicted by PMM. We therefore also divide them into five folds in the experiment.

Note: Our model could perform over the entire dataset but to guarantee the fairness, we only conduct experiment over the subset of users in which PMM could perform in both datasets.

Baselines: We consider several baselines below in this task.

- Center of the mass *COM*: This model returns the center of the mass of all check-ins of a user as his/her home location.
- Most check-in venue *MCV*: This model selects the most frequent check-in venue of a user as his/her home location.
- Periodic Mixture Model *PMM* [5]: It groups check-ins of a user into two clusters named *home* and *work*. The *home* cluster represents non-working hours check-ins and the center of this cluster is the predicted home location of the user.

The two simple baselines *COM* and *MCV* are used for comparison because they appear in previous research works [5, 7, 13].

Performance Measure: We measure the *distance* between the predicted home venue p_i and the actual home location h_i of user i. The overall performance is thus defined by the **average error** ($error_m$) between all predicted home locations and actual home locations. Moreover, we define another metric $prec@k$ is ratio of users whose distance from their predicted home location to actual home is less than k. Formally, $error_m = \frac{\sum_{i\in U} dist(p_i,h_i)}{|U|}$; $prec@k = \frac{|\{i:dist(p_i,h_i)<k\}|}{|U|}$ where $dist(\cdot,\cdot)$ returns the physical distance between two locations. In our experiment, we choose $k = 5km$.

Table 3: Home prediction result of H_SG and H_JK. The unit of average error in this table is meter. The best result of each dataset is highlighted.

		Average Error($prec@5km$)	
	s	**H_SG**	**H_JK**
COM	-	6570.3 (46.2%)	5564.4 (43.4%)
MCV	-	7117.7 (40.3%)	5547.2 (45.5%)
PMM	-	6126.3 (49.3%)	4823.2 (60.8%)
$VAN_{Sigmoid}$	0.1	5561.8 (50.7%)	5623.8 (53.3%)
	0.05	**5046.4** (59.8%)	5125.2 (60.4%)
	0.025	5475.2 (56.7%)	4757.8 (64.4%)
VAN_{CDF}	0.1	5564.6 (51.46%)	5331.1 (56.1%)
	0.05	5181.6 (**60.4%**)	4866.1 (59.1%)
	0.025	5213.8 (56.9%)	**4357.2** (**68.2%**)

Result: Table 3 depicts the performance of baselines and our models with different s parameter values in **H_SG** and **H_JK**.

In the case of **H_SG** dataset, our **VAN**$_{Sigmoid}$ and **VAN**$_{CDF}$ model outperform *PMM* model by 12.34% and 13.16% in term of average error, respectively. Compared with other baselines, the VAN models yield accuracy of average error with up to 28% improvement. Both variants of VAN model also perform better in $prec@5km$ metric. The superior performance of VAN models is not affected by the s parameter.

For **H_JK**, we observe that the performance of our VAN models is affected by the s parameter setting. The optimal s value is 0.025. Under this setting, our VAN models outperform PMM and other baselines. The reason for the poorer performance of other settings may be due to the sparsity of check-ins in this dataset.

5.2 Check-in prediction task

In this section, we evaluate our model in check-in prediction task. This task predicts check-ins between users and venues.

Setup: We sort check-ins in the **H_SG** and **H_JK** datasets by time and then divide each dataset into 10 folds. For each run of experiment, we hide one fold as test set and use the remaining nine folds as training set.

Baselines: For comparison, we use some baselines below

- Probabilistic Matrix Factorization *PMF*[12]: It factorizes check-in matrix into user-latent factor and venue-latent factor matrix alone. We use the number of latent factors $K = 10$.

Table 4: The $recall@k$ of H_SG and H_JK datasets in check-in prediction task. We highlight the best result for each value of k.

		VAN_{CDF}			$VAN_{Sigmoid}$			PMF	MGM	PMF-MGM	N-MF		Expo-MF
	k \ s	0.1	0.05	0.025	0.1	0.05	0.025				100m	200m	
H_SG	20	**4.4%**	1.9%	1.1%	4.2 %	1.95%	0.48%	1.4%	0.36%	1.35%	0.3%	0.29%	1.5%
	50	**8.7%**	4.2%	3.1%	8.67%	4.62%	1.67%	2.6%	0.61%	2.5 %	0.8%	0.75%	1.6%
	100	12.1%	6.6%	6 %	**12.26%**	10.53%	5.65%	3.8%	1.11%	3.7 %	1.4%	1.35%	2.3%
H_JK	20	0.38%	0.56%	0.8%	0.36%	0.74%	**0.96%**	0.14%	0.24%	0.14%	0.57%	0.7%	0.3%
	50	**2.2%**	1.9 %	1.6%	1.73%	1.74%	1.51%	0.29%	0.7 %	0.3 %	0.96%	1.2%	1.1%
	100	**4.3%**	4 %	3.2%	3.8%	3.49%	3 %	1.12%	1.8 %	1.1 %	1.4%	1.7%	2.5%

- Multi-center Gaussian Model *MGM*[4]: It proposed a check-in prediction method using multiple Gaussian distributions as the activity centers of users. We automatically detect the clusters of *MGM* by applying the non-parametric method from Blei *et. al.* [2]. The α parameter of *MGM* which controls the impact of high frequent check-ins venues is set to default value $\alpha = 0.2$.

- Fusion Framework *PMF-MGM*[4]: It combines matrix factorization and *MGM*. It is reported to outperform *PMF* and *MGM* models. We implemented this combined method using *PMF* and *MGM* as its components.

- Matrix Factorization with Neighborhood Influence *N-MF*[8]: It studies the characteristics of geographical neighbors based on the matrix factorization framework. We use the number of latent features $K = 20$ and two venues are neighbors if their distance is less then a predefined threshold d. In our experiment, we set d to 100 meters and 200 meters.

- Exposure Matrix Factorization *Expo-MF*[11]: It incorporates the location of venues and user exposure to increase the performance of check-in prediction under matrix factorization framework. Similar to their experiment, we apply K-Means to cluster venues, the location vector of each venue is its probability to each cluster. We use $K = 100$ for the number of latent factors and the number of clusters in K-Means.

We have different values of latent factors K for *PMF*, *N-MF* and *Expo-MF* because we choose K that produces the best prediction performance for each model in the pool of values 10, 20, 100. These values are default setting reported in the original papers.

Performance Measure: After training, for each user, we select the top k venues predicted by each method and compare against all the venues checked in by the users in the test data.We use $recall@k$ as the metric to compare the performance of our model and the baselines. Finally, we report the average $recall@k$ over all folds. We do not use $precision@k$ because we cannot distinguish between a user disliking a venue and a user not knowing the venue. Formally, $recall@k = \frac{1}{|U|}\sum_{u \in U} \frac{|\mathcal{L}(u,k) \cap \mathcal{L}^t(u)|}{|\mathcal{L}^t(u)|}$ where $\mathcal{L}(u,k)$ is the top k venues of each user u of each predictive method; $\mathcal{L}^t(u)$ represents set of venues of user u in test set and $|\cdot|$ returns the number of elements of set.

Result: The result of check-in prediction task for two datasets **H_SG** and **H_JK** are shown in Table 4 . In our experiment, our model with Sigmoid or CDF function always outperforms all baselines in both datasets. For instance, in **H_SG**, our model could achieve recall up to three times better than PMF and 10 times better than MGM. Overall, in both datasets, if we reduce the area width, the performance of VAN model decreases. Specifically, the performance of area width of 0.05 is usually better than the one of area width of 0.025 but worse than that of area width 0.1. Additionally, the result of VAN_{CDF} usually outperforms $VAN_{Sigmoid}$. *Expo-MF* only considers the location of venues while *N-MF* uses the information of neighbor alone. However, these two baselines do not enjoy better result than both variants of VAN model. Thus, the results suggest that neighborhood competition and area attractiveness are useful factors to be modeled. Between two baselines, *MGM* performans better than *PMF* in **H_JK** dataset but not in **H_SG**. *PMF-MGM* is the hybrid of *MGM* and *PMF* so its performance is in the middle of both models.

6. CONCLUSION

In this paper, we model user visitation by using two new factors: neighborhood competition and area attractiveness. Our experiments show that our model outperforms several baselines in *check-in prediction* and *home location prediction* tasks.

7. ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre@ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

8. REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, 2010.

[2] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 2006.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[4] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, volume 12, 2012.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. KDD, 2011.

[6] T. Doan, F. C. T. Chua, and E. Lim. Mining business competitiveness from user visitation data. In *SBP*, 2015.

[7] T. Doan, F. C. T. Chua, and E. Lim. On neighborhood effects in location-based social networks. In *WI-IAT*, 2015.

[8] L. Hu, A. Sun, and Y. Liu. Your neighbors affect your ratings: On geographical neighborhood influence to rating prediction. In *SIGIR*, 2014.

[9] D. L. Huff. A probabilistic analysis of shopping center trade areas. *Land Economics*, pages 81–90, 1963.

[10] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*, 2012.

[11] D. Liang, L. Charlin, J. McInerney, and D. M. Blei. Modeling user exposure in recommendation. In *WWW*, 2016.

[12] A. Mnih and R. Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, 2007.

[13] Y. Qu and J. Zhang. Trade area analysis using user generated mobile location data. In *WWW*, 2013.

[14] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2012.